Zoea

Research note 36

Building LLMs without Neural Networks

Edward McDaid & Sarah McDaid
1 Jan 2025

Is it possible to build the functional equivalent of a large language model without using neural networks? There is no compelling reason why not and there could be some useful benefits.

Part of the mystique of large language models (LLMs) is the widespread belief that nobody really understands how they actually work. This narrative is even promulgated by some leading figures associated with the technology. Everybody loves a mystery but this could simply be some clever marketing. However, it might just as well be a calculated distraction. This technology is highly investable and paying too much attention to how LLMs function risks trivialising them.

In black box terms, what LLMs do is easy to describe. An LLM is simply an enormous model produced from a large set of documents which captures the probabilities of all the words that might follow any given word in a variety of contexts. Using this model, we can plausibly predict the next word in a partial sentence, and do so repeatedly until the sentence is complete. If, instead of a sentence, we pose a question or state a problem, then the LLM will often produce a reasonable answer or solution. This can include the ability to create other forms of content such as code, if the training set included appropriate examples. Everything that LLMs do is based on this - easy to describe - underlying text generation capability.

There is much debate about the extent to which LLMs simply memorise and regurgitate the information on which they have been trained, like a sort of fuzzy database. Indeed, disputes abound over the legality and damage associated with the wholesale exploitation of copyrighted works in training LLMs. It remains to be seen whether the fair use defence

can be stretched sufficiently to cover the totality of human intellectual output. In any event, nobody denies that LLMs ingest and can recall at least a significant proportion of their training set. In the absence of any real reasoning capabilities, the apparent integration and generalisation of knowledge are merely side effects of the context and recall mechanisms.

One major concern with LLMs is the lack of any clear future development roadmap. Of course, there will be numerous incremental improvements that can be made but there is no obvious next big thing. This means that LLMs look like a technological dead end. As a result, the incessant focus is rather on scaling up - with ever larger models hopefully leading to more impressive results. That strategy might work for a while. However, the power and resource requirements of LLMs are already excessive. According to some estimates, AI probably already exceeds crypto in terms of its global energy usage. Even this gets us nowhere near simulating a single human brain - running at 20 watts.

Excessive power and resource requirements are a direct result of the fact that LLMs are built using neural networks. This is because - for over a generation - neural networks have become the de rigueur technology for all AI problems - a sort of technological monoculture. Neural networks certainly have their uses but they are not necessarily the best approach for all problems. Unfortunately, they are the only technology with which the vast majority of contemporary AI practitioners are familiar. Computer scientists will recognise this as an extreme case of the golden hammer anti-pattern. However, alternatives do exist. For example, most problems that can be solved using neural networks could also be implemented with much greater energy efficiency using Tsetlin machines - a kind of learning automata - or (unfashionably) directly as code.

One of the biggest and most expensive problems for LLMs is the creation of models. This can be a time consuming exercise and a huge investment which must be repeated whenever the training set changes. However, for every LLM that is ever built, probably only a tiny fraction of the model is ever actually used to produce anything. The rest is simply a lot of other people's money that has been turned into heat.

So what if we moved away from the idea of having a huge, up-front, monolithic model and instead used the training set more directly? After all, people are able to integrate and utilise information they have just acquired without having to stop the world and sleep on it first.

It is certainly possible to create software that - at least externally - behaves in exactly the same way as an LLM but which takes the required information directly from the training set. To work, this would need to be able to locate every occurrence of a given word in the training set, and extract the associated following words and preceding context windows. A full text index of the training set could provide all of this information, more or less instantly. An index of this kind would even be smaller than the training set.

Turning this extracted information into next word hypotheses is also fairly straightforward. Indeed, Zoea does something very similar with almost every program it generates. Conceptually, this involves inducing any relationships that might exist between words in the context window and following words in the extract. These relationships can be thought of as a micro-model that is highly specific to the current word and which is produced just-in-time. This in turn directly provides the word hypotheses and probabilities.

So far, this all works in theory but performance might be a concern. If that's the case then Zoea's unique family of powerful heuristics could be used to speed things up dramatically. While they were originally developed to produce code, our heuristics are founded in generic set and language formalisms. This means they will also work with any type of text that has a Zipfian distribution, such as human languages.

So it seems feasible that we could produce something which externally looks and behaves like an LLM yet requires no training, runs on modest infrastructure and uses comparatively little energy. Furthermore, the training set can be increased or even decreased at any time and instantly reflected in the results. All processing is also transparent and comprehensible. These are some compelling benefits, unless you happen to be a data-centre operator, GPU manufacturer, power company or LLM investor.

Aside from saving the planet there would be a number of other benefits to this exercise. The problem of LLMs being a dead end is exacerbated by the belief that nobody understands how they work. Replacing LLMs with a technology that anyone can understand should improve the prospects for future innovation. The use of neural networks and other arcane LLM componentry makes it difficult to understand how information is processed. This is at odds with the transparent, unbiased and even legal use of AI that most people really want. Finally, AI needs more diversity in terms of the technological

approaches and tools upon which it depends. While neural networks are undoubtedly useful, there are problems for which they aren't the right solution. We need to be more open minded about our technology choices in such cases.

There is also an interesting implication here that maybe model building should to be an integral part of constructing and running AI solutions - not a separate and disconnected activity. Currently, AI models are built speculatively, in advance and wholesale. The high resource requirements for model building encourage a batch oriented mentality, leading to models that are static rather than dynamic. This seems to be accepted by everyone as a given but it doesn't have to be so.

So what is the point of all this? After all, Zoea does not currently use LLMs, or even neural networks for that matter. However, if we ever did need something like an LLM then this is how we would build it. A more immediate consideration is the current level of hype surrounding AI in general and LLMs in particular. Failure to deliver on an over inflated level of expectations risks triggering another AI winter. This would be bad news for the whole AI industry. Alternative strategies need to be available.

LLMs are certainly an important achievement but they represent a local maxima in AI solution space. The next big thing is somewhere completely different but, as the saying goes, you can't get there from here.

Learn more at **zoea.co.uk**